D36–D42 *Nucleic Acids Research, 2013, Vol. 41, Database issue* *Published online 27 November 2012*
doi:10.1093/nar/gks1195

GenBank

Dennis A. Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell and Eric W. Sayers*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA

Received September 28, 2012; Revised and Accepted October 29, 2012

ABSTRACT

GenBank® (http://www.ncbi.nlm.nih.gov) is a comprehensive database that contains publicly available nucleotide sequences for almost 260 000 formally described species. These sequences are obtained primarily through submissions from individual laboratories and batch submissions from large-scale sequencing projects, including whole-genome shotgun (WGS) and environmental sampling projects. Most submissions are made using the web-based BankIt or standalone Sequin programs, and GenBank staff assigns accession numbers upon data receipt. Daily data exchange with the European Nucleotide Archive (ENA) and the DNA Data Bank of Japan (DDBJ) ensures worldwide coverage. GenBank is accessible through the NCBI Entrez retrieval system, which integrates data from the major DNA and protein sequence databases along with taxonomy, genome, mapping, protein structure and domain information, and the biomedical journal literature via PubMed. BLAST provides sequence similarity searches of GenBank and other sequence databases. Complete bimonthly releases and daily updates of the GenBank database are available by FTP. To access GenBank and its related retrieval and analysis services, begin at the NCBI home page: www.ncbi.nlm.nih.gov.

INTRODUCTION

GenBank (1) is a comprehensive public database of nucleotide sequences and supporting bibliographic and biological annotation. GenBank is built and distributed by the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM), located at the campus of the U.S. National Institutes of Health (NIH) in Bethesda, MD, USA.

NCBI builds GenBank primarily from the submission of sequence data from authors and from the bulk submission of expressed sequence tag (EST), genome survey sequence (GSS), whole-genome shotgun (WGS) and other high-throughput data from sequencing centres. The U.S. Patent and Trademark Office also contributes sequences from issued patents. GenBank participates with the European Molecular Biology Laboratory Nucleotide Sequence Database (EMBL-Bank), part of the European Nucleotide Archive (ENA) (2), and the DNA Data Bank of Japan (DDBJ) (3) as a partner in the International Nucleotide Sequence Database Collaboration (INSDC). The INSDC partners exchange data daily to ensure that a uniform and comprehensive collection of sequence information is available worldwide. NCBI makes the GenBank data available at no cost over the Internet, through FTP and a wide range of web-based retrieval and analysis services (4).

RECENT DEVELOPMENTS

Submission portal

NCBI is in the process of creating a unified submission portal that will provide a single access point for data submitters (submit.ncbi.nlm.nih.gov). Submitters will be able to create accounts that will track and display all of their submissions and will facilitate communication with relevant NCBI staff. With respect to GenBank, the portal now supports submissions of whole genome shotgun (WGS) and transcriptome shotgun assembly (TSA) sequences and, in the near future, complete microbial genomes. Submitters may continue to use standard GenBank submission tools (see below) for other GenBank submissions.

New submission wizards

The Sequin program, a popular tool for preparing electronic submissions (see below), now contains a variety of 'wizards' to assist users when submitting particular types of sequences. The current release of Sequin (version 12.21) contains wizards for submitting viral sequences; uncultured sequences; rRNA, internal transcribed spacer and rRNA-intergenic spacer sequences (rRNA-ITS-IGS); TSA sequences and non-rRNA intergenic spacer sequences (IGS). The BankIt web submission tool also has a special

*To whom correspondence should be addressed. Tel: +1 301 496 2475; Fax: +1 301 480 9241; Email: sayers@ncbi.nlm.nih.gov

function to assist with submitting 16S rRNA sequences. All of these tools guide users through the process of submitting the required data and also assist with sequence annotation.

WGS browser

Within GenBank, WGS master records (see below) contain no sequence data, but rather show the descriptive information and range of accession numbers of the contigs submitted as part of that WGS project. NCBI is transitioning to a point where we will no longer assign GI (GenInfo) numbers to these individual contigs, particularly for data from low coverage, fragmented or unannotated assemblies of eukaryotic genomes. Contigs without GI numbers will not be available from the Nucleotide database; instead, users may view these records in the WGS browser linked from the WGS feature of any WGS master record. The WGS browser provides the complete descriptive information from the master record of the project, interactive views of the FASTA of every contig record and also provides links to the FTP files for all the contigs of the entire project.

New TSA accessions

NCBI is now formatting and releasing TSA records similarly to what has been done for WGS data (see below). Like WGS, TSA projects will now contain a master record, in addition to records representing each of the assembled contigs. TSA will be using a similar accession number scheme to WGS as well. Like WGS accessions, the new TSA accessions have a four-letter prefix, representing the TSA project, followed by a two-digit version number and a six-digit contig number. For example, GAAA01000020 is contig 20 from the first version of TSA project GAAA. In the future, the individual TSA contigs will not be indexed in Entrez, but will be available in the WGS browser. The TSA project master records have accessions that begin with the four-letter prefix followed by eight zeroes (e.g. GAAA00000000) and are indexed in the Nucleotide database.

Both TSA and WGS records will now also contain an Assembly block in the COMMENT section of their GenBank reports. The Assembly block contains, as available, information about the assembly method, the assembly name, the genome coverage achieved and the sequencing technology used to generate the data. A sample Assembly block from GAAA00000000 is shown below:

```
##Assembly-Data-START##
Assembly Method         :: Trinity v. r2011-07-13
Assembly Name           :: LatCha_Muscle767971_v1.0
Coverage                :: 590x
Sequencing Technology   :: Illumina Hi-Seq
##Assembly-Data-END##
```

ORGANIZATION OF THE DATABASE

GenBank divisions

GenBank assigns sequence records to various divisions based either on the source taxonomy or the sequencing strategy used to obtain the data. There are 12 'taxonomic' divisions that correspond roughly to the source organisms of the sequence data (BCT, ENV, INV, MAM, PHG, PLN, PRI, ROD, SYN, UNA, VRL, VRT) and 8 'functional' divisions (EST, GSS, HTC, HTG, PAT, STS, TSA, WGS) that collect sequences generated by a particular method. The size and growth of these divisions, and of GenBank as a whole, are shown in Table 1.

Sequence-based taxonomy

Database sequences are classified and can be queried using a comprehensive sequence-based taxonomy (www.ncbi.nlm.nih.gov/taxonomy/) developed by NCBI in collaboration with EMBL-Bank and DDBJ and with the valuable assistance of external advisers and curators (5). Almost 260 000 formally described species are represented in GenBank, and the top species in the non-WGS GenBank divisions are listed in Table 2.

Sequence identifiers and accession numbers

Each GenBank record, consisting of both a sequence and its annotations, is assigned a unique identifier called an accession number that is shared across the three collaborating databases (GenBank, DDBJ, EMBL-Bank). The accession number appears on the ACCESSION line of a GenBank record and remains constant over the lifetime of the record, even when there is a change to the sequence or annotation. Changes to the sequence data are tracked by an integer extension of the accession number, and this *Accession.version* identifier appears on the VERSION line of the GenBank flat file. Other changes, such as revised annotations or additions of publications, that do not affect the sequence data will not result in a new version number. The initial version of a sequence has the extension '.1'. In addition, each version of the DNA sequence is also assigned a unique NCBI identifier called a GI number that also appears on the VERSION line following the *Accession.version*:

```
ACCESSION AF000001
VERSION AF000001.5 GI: 7274584
```

Each GI number corresponds to a unique *Accession.version* identifier. When a change is made to a sequence in a GenBank record, a new GI number is issued to the updated sequence and the version extension of the *Accession.version* identifier is incremented. The accession number for the record as a whole remains unchanged, and will always retrieve the most recent version of the record; the older versions remain available under the old *Accession.version* identifiers and their original GI numbers. The Revision History report, available from the 'Display Settings' menu on the sequence record view, summarizes the various updates for that GenBank record, both those that resulted in a new version (updates to sequence data) and those that did not (updates to non-sequence data).

A similar system tracks changes in the corresponding protein translations. These identifiers appear as qualifiers for coding sequence (CDS) features in the FEATURES portion of a GenBank entry, e.g. /protein_id = 'AAF14809.1'. Protein sequence translations also receive their own

Table 1. Growth of GenBank divisions (nucleotide base pairs)

Division	Description	Release 191 (8/2012)	Annual increase (%)[a]
Taxnomic divisons			
SYN	Synthetic	928 200 038	494.2%
PHG	Phages	84 079 451	34.4%
ENV	Environmental samples	3 374 433 548	32.1%
VRL	Viruses	1 429 464 786	21.1%
BCT	Bacteria	8 439 854 434	21.0%
PLN	Plants	5 481 470 133	15.6%
MAM	Other mammals	863 036 872	6.9%
VRT	Other vertebrates	2 886 594 595	6.7%
PRI	Primates	6 317 656 773	3.3%
UNA	Unannotated	127 803	1.5%
ROD	Rodents	4 435 106 948	0.9%
INV	Invertebrates	2 493 058 927	−1.7%
Functional divisions			
TSA	Transcriptome shotgun data	5 759 588 580	207.3%
WGS	Whole-genome shotgun data	308 196 411 905	47.9%
PAT	Patented sequences	12 118 622 726	8.6%
GSS	Genome survey sequences	21 947 780 105	5.7%
EST	Expressed sequence tags	40 888 051 100	4.8%
HTG	High-throughput genomic	24 359 210 558	0.1%
STS	Sequence tagged sites	636 262 446	0.1%
HTC	High-throughput cDNA	639 165 410	−3.5%
TOTAL	All GenBank sequences	451 278 177 138	33.1%

[a]Measured relative to Release 185 (8/2011).

Table 2. Top organisms in GenBank (Release 191)

Organism	Non-WGS base pairs
Homo sapiens	16 310 774 187
Mus musculus	9 974 977 889
Rattus norvegicus	6 521 253 272
Bos taurus	5 386 258 455
Zea mays	5 062 731 057
Sus scrofa	4 887 861 860
Danio rerio	3 120 857 462
Strongylocentrotus purpuratus	1 435 236 534
Macaca mulatta	1 256 203 101
Oryza sativa Japonica Group	1 255 686 573
Xenopus (Silurana) tropicalis	1 249 938 611
Nicotiana tabacum	1 197 357 811
Arabidopsis thaliana	1 144 226 616
Drosophila melanogaster	1 119 965 220
Pan troglodytes	1 008 323 292
Vitis vinifera	999 010 073
Canis lupus familiaris	951 238 343
Glycine max	906 638 854
Gallus gallus	899 631 338
Triticum aestivum	898 689 329

unique GI number, which appears as a second qualifier on the CDS feature:

```
/db_xref = 'GI:6513858'
```

Citing GenBank records

Besides being the primary identifier of a GenBank sequence record, GenBank accessions are also the most efficient and reliable way to cite a sequence record in publications. We certainly encourage submitters and other authors to cite GenBank data using these accessions. However, as discussed above, since searching with a GenBank accession number will retrieve the most recent version of the sequence data for a record, the sequence data returned from such searches will change over time if the record is updated. It is quite possible, therefore, for the sequence data retrieved today by an accession to be different from that discussed or analysed in an article published several years ago. We therefore recommend that authors include the version suffix when citing a GenBank accession (e.g. AF000001.5), particularly in cases where the sequence coordinates are critical to the work being described.

BUILDING THE DATABASE

The data in GenBank and the collaborating databases, EMBL-Bank and DDBJ, are submitted either by individual authors to one of the three databases or by sequencing centres as batches of EST, STS, GSS, HTC, TSA, WGS or HTG sequences. Data are exchanged daily with DDBJ and EMBL-Bank so that the daily updates from NCBI servers incorporate the most recently available sequence data from all sources.

Direct electronic submission

Virtually all records enter GenBank as direct electronic submissions (www.ncbi.nlm.nih.gov/genbank/), with the majority of authors using the BankIt or Sequin programs. An online table (www.ncbi.nlm.nih.gov/guide/howto/submit-sequence-data/) provides general guidance and links to appropriate tools for submitting a variety of sequence data. Many journals require authors with sequence data to submit the data to a public sequence database as a condition of publication. GenBank staff can usually assign an accession number to a sequence

submission within two working days of receipt, and do so at a rate of ~3500 per day. The accession number serves as confirmation that the sequence has been submitted and provides a means for readers of articles in which the sequence is cited to retrieve the data. Direct submissions receive a quality assurance review that includes checks for vector contamination, proper translation of coding regions, correct taxonomy and correct bibliographic citations. A draft of the GenBank record is passed back to the author for review before it enters the database.

Authors may ask that their sequences be kept confidential until the time of publication. Since GenBank policy requires that the deposited sequence data be made public when the sequence or accession number is published, authors are instructed to inform GenBank staff of the publication date of the article in which the sequence is cited to ensure a timely release of the data. Although only the submitter is permitted to modify sequence data or annotations, all users are encouraged to report lags in releasing data or possible errors or omissions to GenBank at update@ncbi.nlm.nih.gov.

NCBI works closely with sequencing centres to ensure timely incorporation of bulk data into GenBank for public release. GenBank offers special batch procedures for large-scale sequencing groups to facilitate data submission, including the program *tbl2asn*, described at www.ncbi.nlm.nih.gov/genbank/tbl2asn2.html.

Submission using BankIt

About a third of author submissions are received through an NCBI web-based data submission tool named BankIt. Using BankIt, authors enter sequence information and biological annotations, such as coding regions or mRNA features, directly into a series of tabbed forms that allow the submitter to describe the sequence further without having to learn formatting rules or controlled vocabularies. Additionally, BankIt allows submitters to upload source and annotation data using tab-delimited tables. Before creating a draft record in the GenBank flat file format for the submitter to review, BankIt validates the submissions by flagging many common errors and checking for vector contamination using a variant of BLAST called Vecscreen.

Submission using Sequin and tbl2asn

NCBI also offers a standalone multi-platform submission program called Sequin (www.ncbi.nlm.nih.gov/projects/Sequin/) that can be used interactively with other NCBI sequence retrieval and analysis tools. Sequin handles simple sequences (such as a single cDNA), phylogenetic studies, population studies, mutation studies, environmental samples with or without alignments and sequences with complex annotation. Sequin has convenient editing and complex annotation capabilities and contains a number of built-in validation functions for quality assurance. Sequin is able to accommodate sequences such as the 5.6 Mb *E. coli* genome and read in a full complement of annotations from simple tables. The most recent version, Sequin 12.2, was released in June 2012 and is available for Macintosh, PC and Unix computers via anonymous FTP at ftp.ncbi.nlm.nih.gov/sequin. Once a submission is completed, submitters can email the Sequin file to gb-sub@ncbi.nlm.nih.gov. Submitters of large, heavily annotated genomes are encouraged to use the command line tool *tbl2asn* to convert a table of annotations generated from an annotation pipeline into an ASN.1 (Abstract Syntax Notation One) record suitable for submission to GenBank.

Submission of Barcode sequences

The Consortium for the Barcode of Life (CBOL, www.barcoding.si.edu/) is an international initiative to develop DNA barcoding as a tool for characterizing species of organisms using a short DNA sequence. For animal species, a 648-base pair fragment of the gene for cytochrome oxidase subunit I is used as the barcode. The plant and fungal communities are using other loci. NCBI provides an online tool (BarSTool) for the bulk submission of barcode sequences to GenBank (www.ncbi.nlm.nih.gov/WebSub/?tool=barcode) that allows users to upload files containing a batch of sequences with associated source information. Barcode sequences can be retrieved from the Nucleotide database with the query 'barcode[keyword]'.

Additional notes on special divisions and record types

Transcriptome Shotgun Assembly (TSA) sequences

The TSA division contains transcriptome shotgun assembly sequences that are assembled from sequences deposited in the NCBI Trace Archive, the Sequence Read Archive (SRA) and the EST division of GenBank. Although neither the Trace Archive nor SRA is a part of GenBank, they are part of the INSDC and provide access to the data underlying these assemblies (4,6). TSA records have 'TSA' as their keyword and can be retrieved with the query 'tsa[properties]'. TSA continues to be one of the most rapidly growing divisions of GenBank, more than tripling in size over the past year (Table 1).

Environmental sample sequences (ENV)

The ENV division of GenBank accommodates sequences obtained via environmental sampling methods in which the source organism is unknown. Many ENV sequences arise from metagenome samples derived from microbiota in various animal tissues, such as within the gut or skin, or from particular environments, such as freshwater sediment, hot springs or areas of mine drainage. Records in the ENV division contain 'ENV' in the keyword field and use an '/environmental_sample' qualifier in the source feature. Environmental sample sequences are generally submitted for whole metagenomic shotgun sequencing experiments or surveys of sequences from targeted genes, like 16S rRNA. NCBI continues to support BLAST searches (see below) of metagenomic ENV sequences, but sequences within WGS projects are now part of the WGS BLAST database.

Whole-Genome Shotgun sequences

Whole-Genome Shotgun (WGS) sequences appear in GenBank as groups of sequence-overlap contigs collected under a master WGS record. Each master record represents a WGS project and has an accession number in the

Nucleotide database consisting of a four-letter prefix followed by eight zeroes and a version suffix as found in standard GenBank records. The number of zeroes increases to nine for WGS projects with one million or more contigs. Master records contain no sequence data; rather, they are linked to their set of individual contigs that can be viewed using the new WGS browser (see above). Contig records have accessions consisting of the same four-letter prefix as their master accession, followed by a two-digit version number and a six-digit contig ID. For example, the WGS accession number 'AAAA02002744' is assigned to contig number '002744' of the second version of project 'AAAA', whose accession number is 'AAAA00000000.2'. Currently, there are >6000 WGS sequencing projects, many of whose data have been used to build almost 12 million scaffolds and chromosomes for genome assemblies. For a complete list of WGS projects with links to the data, see www.ncbi.nlm.nih.gov/Traces/wgs/.

Although WGS project sequences may be annotated, many low-coverage genome projects do not contain annotation. Because these sequence projects are ongoing and incomplete, these annotations may not be tracked from one assembly version to the next and should be considered preliminary. Submitters of genomic sequences, including WGS sequences, are urged to use evidence tags of the form '/experimental=*CATEGORY:text*' and '/inference=*CATEGORY:TYPE:text*', where *TYPE* is one of a number of standard inference types, *text* consists of structured text and the optional *CATEGORY* label is one of the following:

COORDINATES—support for the annotated coordinates

DESCRIPTION—support for a broad concept of function such as that based on phenotype, genetic approach, biochemical function, pathway information, etc.

EXISTENCE—support for the known or inferred existence of the product.

Expressed sequence tags (ESTs)

ESTs continue to be a major source of data for gene expression and annotation studies, and at almost 41 billion base pairs, it remains the largest non-WGS division in GenBank. EST data are available for download from ftp.ncbi.nlm.nih.gov/repository/dbEST/ (7) as well as from the GenBank FTP site. The data in dbEST are clustered using the BLAST programs to produce the UniGene database (www.ncbi.nlm.nih.gov/unigene) of >5.8 million gene-oriented sequence clusters representing 142 organisms (4).

High-throughput genomic (HTG) and high-throughput cDNA (HTC) sequences

The HTG division of GenBank (www.ncbi.nlm.nih.gov/genbank/htgs/) contains unfinished large-scale genomic records, which are in transition to a finished state (8). These records are designated as belonging to Phases 0 to 3 depending on the quality of the data, with Phase 3 being the finished state. On reaching Phase 3, HTG records are moved into the appropriate organism division of GenBank.

The HTC division of GenBank contains high-throughput cDNA sequences that are of draft quality but may contain 5′ untranslated regions (UTRs), 3′ UTRs, partial coding regions and introns. HTC sequences that are finished and of high quality are moved to the appropriate organism division of GenBank. A project generating HTC data is described in (9).

Third Party Annotation

Third Party Annotation (TPA) records are sequence annotations published by someone other than the original submitter of the primary sequence record in DDBJ/ENA/GenBank (www.ncbi.nlm.nih.gov/genbank/TPA). Each of the current 164 000 TPA records falls into one of three categories: *experimental*, in which case there is direct experimental evidence for the existence of the annotated molecule; *inferential*, in which case the experimental evidence is indirect; and *reassembly*, where the focus is on providing a better assembly of the raw reads. TPA sequences may be created by assembling a number of primary sequences. The format of a TPA record (e.g. BK000016) is similar to that of a conventional GenBank record but includes the label 'TPA_exp:', 'TPA_inf:' or 'TPA_reasm:' at the beginning of each Definition Line as well as corresponding keywords. TPA experimental and inferential records also contain a Primary block that provides the base ranges and identifier for the sequences used to build the TPA. TPA sequences are not released to the public until their accession numbers or sequence data and annotation appear in a peer-reviewed biological journal. TPA submissions to GenBank may be made using either BankIt or Sequin.

Contig (CON) records for assemblies of smaller records

Within GenBank, CON records are used to represent very long sequences, such as a eukaryotic chromosome, where the sequence is not complete but consists of several contig records with uncharacterized gaps between them. Rather than listing the sequence itself, CON records contain assembly instructions involving the several component sequences. An example of such a CON record is CM000663 for human chromosome 1.

RETRIEVING GENBANK DATA

The Entrez system

The sequence records in GenBank are accessible through the NCBI Entrez retrieval system (4). Records from the EST and GSS divisions of GenBank are stored in the EST and GSS databases, whereas all other GenBank records are stored in the Nucleotide database (Table 3). GenBank sequences that are part of population or phylogenetic studies are also collected together in the PopSet database, and conceptual translations of CDS sequences annotated on GenBank records are available in the Protein database. Each of these databases is linked to the scientific literature in PubMed and PubMed Central. Additional information about conducting Entrez searches is found in the NCBI Help Manual (www.ncbi.nlm.nih.gov/books/NBK3831/) and links to related tutorials are provided on the NCBI Education page (www.ncbi.nlm.nih.gov/education/).

Table 3. Retrieval databases containing GenBank data

Division	Entrez database	BLAST database
BCT, ENV, INV, MAM, PHG, PLN, PRI, ROD, SYN, UNA, VRL, VRT	nucleotide	nr
EST	est	est
GSS	gss	gss
HTC	nucleotide	nr
HTG	nucleotide	htg
PAT	nucleotide	pat
STS	nucleotide	dbsts
TSA	nucleotide	tsa
WGS	nucleotide	wgs

Associating sequence records with sequencing projects

The ability to identify all GenBank records submitted by a specific group or those with a particular focus, such as metagenomic surveys, is essential for the analysis of large volumes of sequence data. The use of organism or submitter names as a means to define such a set of sequences is unreliable. The BioProject database (www.ncbi.nlm.nih.gov/bioproject), developed at NCBI and subsequently adopted across the INSDC, allows submitters to register large-scale sequencing projects under a unique project identifier, enabling reliable linkage between sequencing projects and the data they produce (10). BioProject includes pointers to data from a wide variety of projects deposited in any NCBI primary data archive. Sequencing projects focus on genomes, metagenomes, transcriptomes, comparative genomics as well as on particular loci, such as 16S ribosomal RNA. A 'DBLINK' line appearing in GenBank flat files identifies the sequencing projects with which a GenBank sequence record is associated. In addition, sequence records may now have a link to the BioSample database (10) that provides additional information about the biological materials used in the study that produced the sequence data. Such studies include genome-wide association studies, high-throughput sequencing, microarrays and epigenomic analyses. As an example, the TSA project GAAA (see above) contains DBLINK lines that associate the GenBank sequence record with BioProject record PRJNA54005 and BioSample record SRS283232, as well as the SRA record containing the raw data, SRR401852:

```
BioProject: PRJNA77699
BioSample: SRS283232
Sequence Read Archive: SRR401852
```

Another example is the Human Microbiome Project (HMP) that is represented by the umbrella BioProject 43021 (www.ncbi.nlm.nih.gov/bioproject/43021). Users can then find sequence data by following links to the various subprojects listed on this record.

BLAST sequence-similarity searching

Sequence-similarity searches are the most fundamental and frequent type of analysis performed on GenBank data. NCBI offers the BLAST family of programs (blast.ncbi.nlm.nih.gov) to detect similarities between a query sequence and database sequences (11,12). BLAST searches may be performed on the NCBI website (13) or by using a set of standalone programs distributed by FTP (4). Table 3 displays the appropriate BLAST databases for the various divisions of GenBank.

Obtaining GenBank by FTP

NCBI distributes GenBank releases in the traditional flat file format as well as in the ASN.1 format used for internal maintenance. The full bimonthly GenBank release along with the daily updates, which incorporate sequence data from EMBL-Bank and DDBJ, is available by anonymous FTP from NCBI at ftp.ncbi.nlm.nih.gov/genbank. GenBank is also available for high-speed download using an Aspera client at www.ncbi.nlm.nih.gov/public/. The full release in flat file format is available as a set of compressed files with a non-cumulative set of updates at ftp.ncbi.nlm.nih.gov/genbank/daily-nc/. For convenience in file transfer, the data are partitioned into multiple files; for release 191, there are 1852 files requiring 604 GB of uncompressed disk storage. A script is provided in ftp.ncbi.nlm.nih.gov/genbank/tools/ to convert a set of daily updates into a cumulative update.

FOR MORE INFORMATION

Additional information about GenBank is available on the main GenBank web page (www.ncbi.nlm.nih.gov/genbank) and the Entrez Sequences Help Manual (www.ncbi.nlm.nih.gov/books/NBK44864/). The NCBI Education page (www.ncbi.nlm.nih.gov/Education/) lists links to NCBI documentation, tutorials and educational tools along with links to outreach initiatives including Discovery Workshops, webinars and upcoming conference exhibits. NCBI provides updates to GenBank and other resources by RSS (www.ncbi.nlm.nih.gov/feed/) and on Twitter and Facebook (links are in the common footer of NCBI pages). Users may also want to consult the bionet GenBank newsgroup (www.bio.net/bionet/mm/genbankb/). This newsgroup is not managed by NCBI, but NCBI staff are regular contributors. Finally, a complete description of each GenBank release is provided in the gbrel.txt file distributed as part of the release, and an archive of these files is provided at ftp.ncbi.nlm.nih.gov/genbank/release.notes/.

MAILING ADDRESS

GenBank, National Center for Biotechnology Information, Building 45, Room 6AN12D-37, 45 Center Drive, Bethesda, MD 20892, USA.

ELECTRONIC ADDRESSES

www.ncbi.nlm.nih.gov—NCBI Home Page.

gb-sub@ncbi.nlm.nih.gov—Submission of sequence data to GenBank.

update@ncbi.nlm.nih.gov—Revisions to, or notification of release of, 'confidential' GenBank entries.

info@ncbi.nlm.nih.gov—General information about NCBI resources.

CITING GENBANK

If you use the GenBank database in your published research, we ask that this article be cited.

FUNDING

Funding for open access charge: Intramural Research Program of the National Institutes of Health; National Library of Medicine.

Conflict of interest statement. None declared.

REFERENCES

1. Benson,D.A., Karsch-Mizrachi,I., Clark,K., Lipman,D.J., Ostell,J. and Sayers,E.W. (2012) GenBank. *Nucleic Acids Res.*, **40**, D48–D53.
2. Leinonen,R., Akhtar,R., Birney,E., Bower,L., Cerdeno-Tarraga,A., Cheng,Y., Cleland,I., Faruque,N., Goodgame,N., Gibson,R. *et al.* (2011) The European Nucleotide Archive. *Nucleic Acids Res.*, **39**, D28–D31.
3. Kaminuma,E., Kosuge,T., Kodama,Y., Aono,H., Mashima,J., Gojobori,T., Sugawara,H., Ogasawara,O., Takagi,T., Okubo,K. *et al.* (2011) DDBJ progress report. *Nucleic Acids Res.*, **39**, D22–D27.
4. NCBI Resource Coordinators. (2013) Database resources at the National Center for Biotechnology Information. *Nucleic Acids Res.*, **41**, D8–D20.
5. Federhen,S. (2012) The NCBI Taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.
6. Kodama,Y., Shumway,M. and Leinonen,R. (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.
7. Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) dbEST—database for "expressed sequence tags". *Nat. Genet.*, **4**, 332–333.
8. Kans,J.A. and Ouellette,B.F.F. (2001) Submitting DNA Sequences to the Databases. In: Baxevanis,A.D. and Ouellette,B.F.F. (eds), *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. John Wiley and Sons, Inc., New York, NY, pp. 65–81.
9. Kawai,J., Shinagawa,A., Shibata,K., Yoshino,M., Itoh,M., Ishii,Y., Arakawa,T., Hara,A., Fukunishi,Y., Konno,H. *et al.* (2001) Functional annotation of a full-length mouse cDNA collection. *Nature*, **409**, 685–690.
10. Barrett,T., Clark,K., Gevorgyan,R., Gorelenkov,V., Gribov,E., Karsch-Mizrachi,I., Kimelman,M., Pruitt,K.D., Resenchuk,S., Tatusova,T. *et al.* (2012) BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res.*, **40**, D57–D63.
11. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
12. Zhang,Z., Schaffer,A.A., Miller,W., Madden,T.L., Lipman,D.J., Koonin,E.V. and Altschul,S.F. (1998) Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res.*, **26**, 3986–3990.
13. Johnson,M., Zaretskaya,I., Raytselis,Y., Merezhuk,Y., McGinnis,S. and Madden,T.L. (2008) NCBI BLAST: a better web interface. *Nucleic Acids Res.*, **36**, W5–W9.